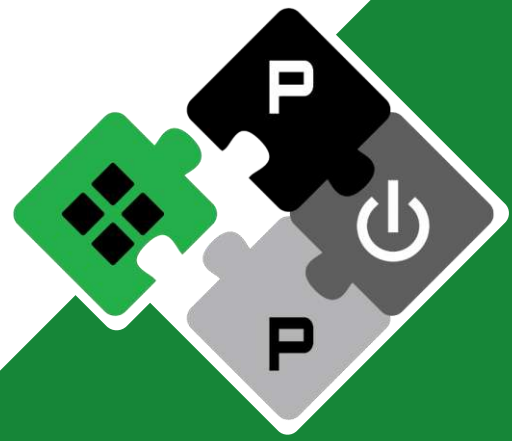
ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Parallel Ultra-Low Power (PULP) Processing for Next-Generation Wearable EEG Monitoring

Prof. Dr. Luca Benini
lbenini@iis.ee.ethz.ch



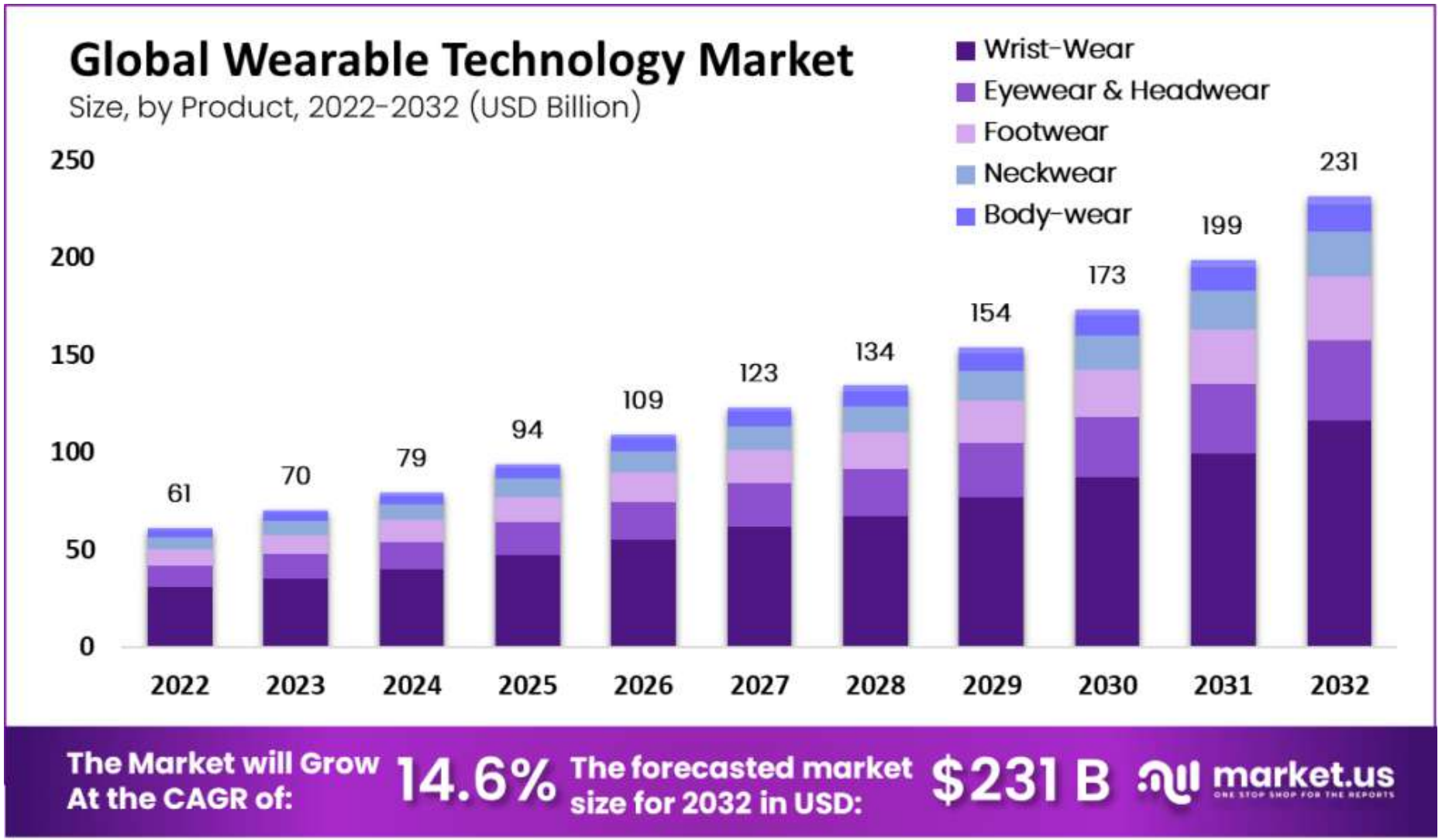
@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

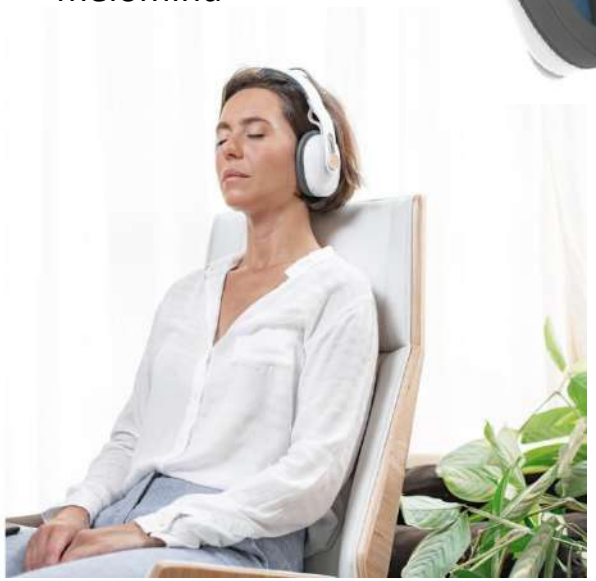
PULP Platform
Open Source Hardware, the way it should be!

Wearable devices



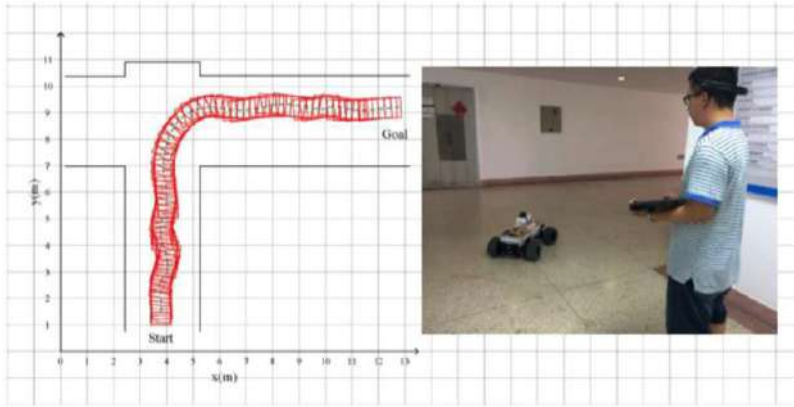
Consumer Wearable EEG devices

melomind



Emotiv

versus



[1]

Ground vehicle control

Muse

[1] J. Zhuang, et al. "Ensemble Learning Based Brain–Computer Interface System for Ground Vehicle Control," in IEEE TSMC: Systems, 2021.

Requirements for a Successful Wearable Device

Safe

Privacy preserving

Comfortable, no stigma

Accurate

Long battery life

Real-time response

Current solutions

Transmission

Data acquisition

Data processing

Feedback to user

Bandwidth

Power hungry

Short lifetime

Privacy

Latency

Next generation wearable devices

Data acquisition

Transmission

Energy efficient edge computing
↓
Smart wearables!

Data processing

Feedback to user

- Massive cost reduction
- Reduced power consumption
- Longer lifetime
- Protection of user data
- Reduced latency

Challenges

Data acquisition

Transmission

Energy efficient edge computing ↓ Smart wearables!

Data processing

Feedback to user

Low power

High performance

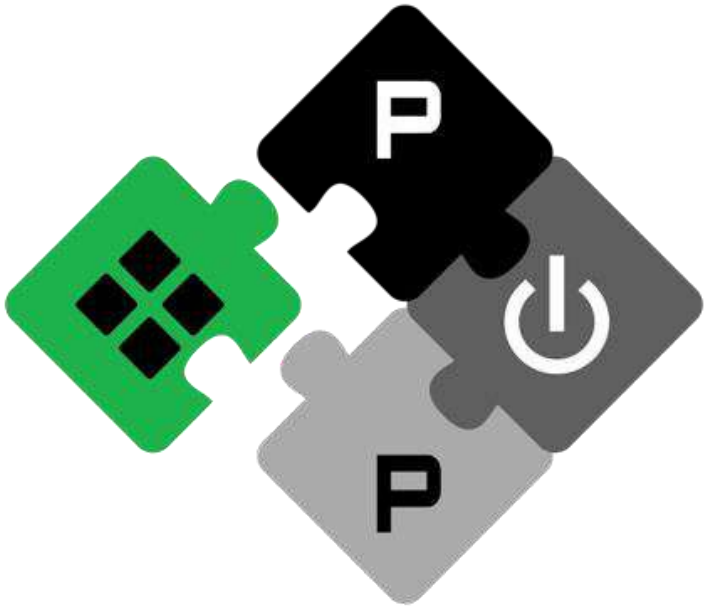
Accurate

Limited resources

Energy efficient deployment

Wearable MCU 100MHz, Mobile SoC 1GHz x 10 cores → Want 100x more compute at same power… How?

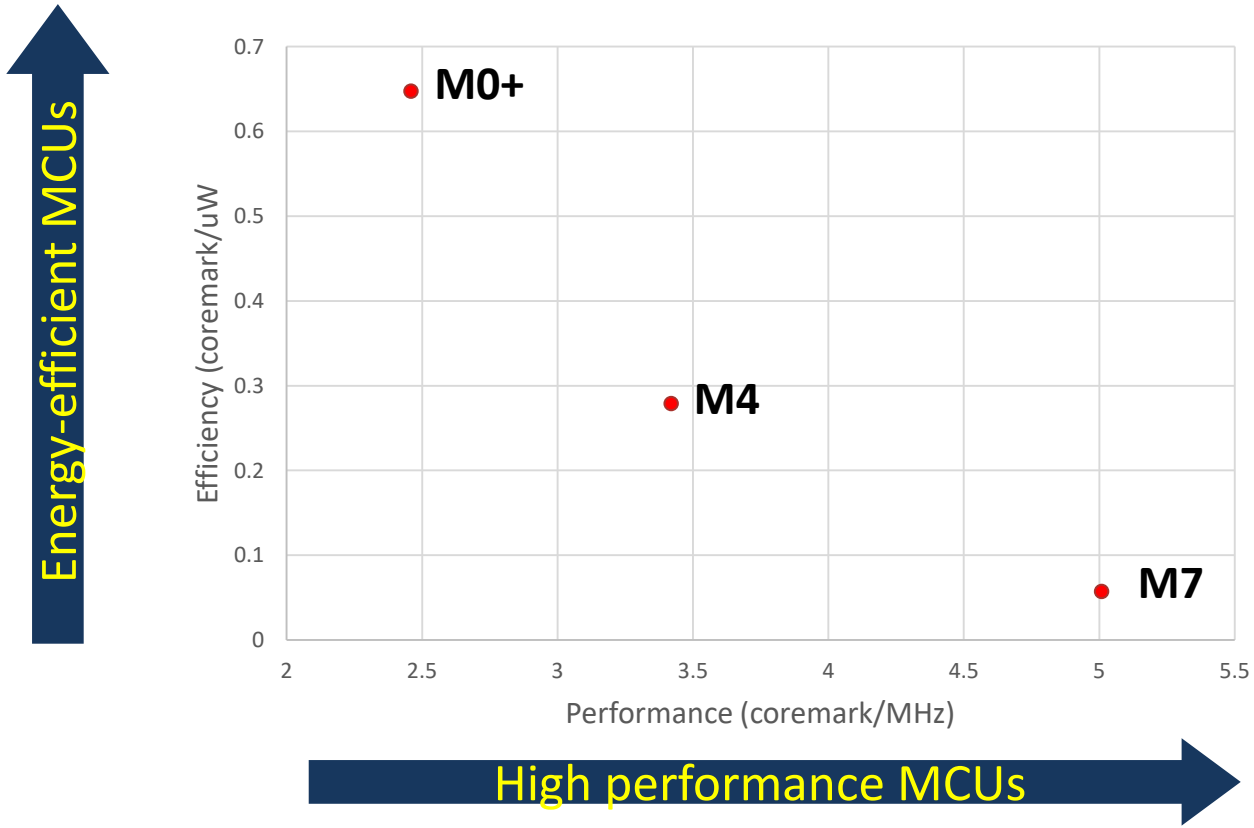
Challenges



Wearable MCU 100MHz, Mobile SoC 1GHz x 10 cores → Want 100x more compute at same power… How?

The Challenge: Energy efficiency@GOPS

ARM Cortex-M MCUs: M0+, M4, M7 (40LP, typ, 1.1V)*

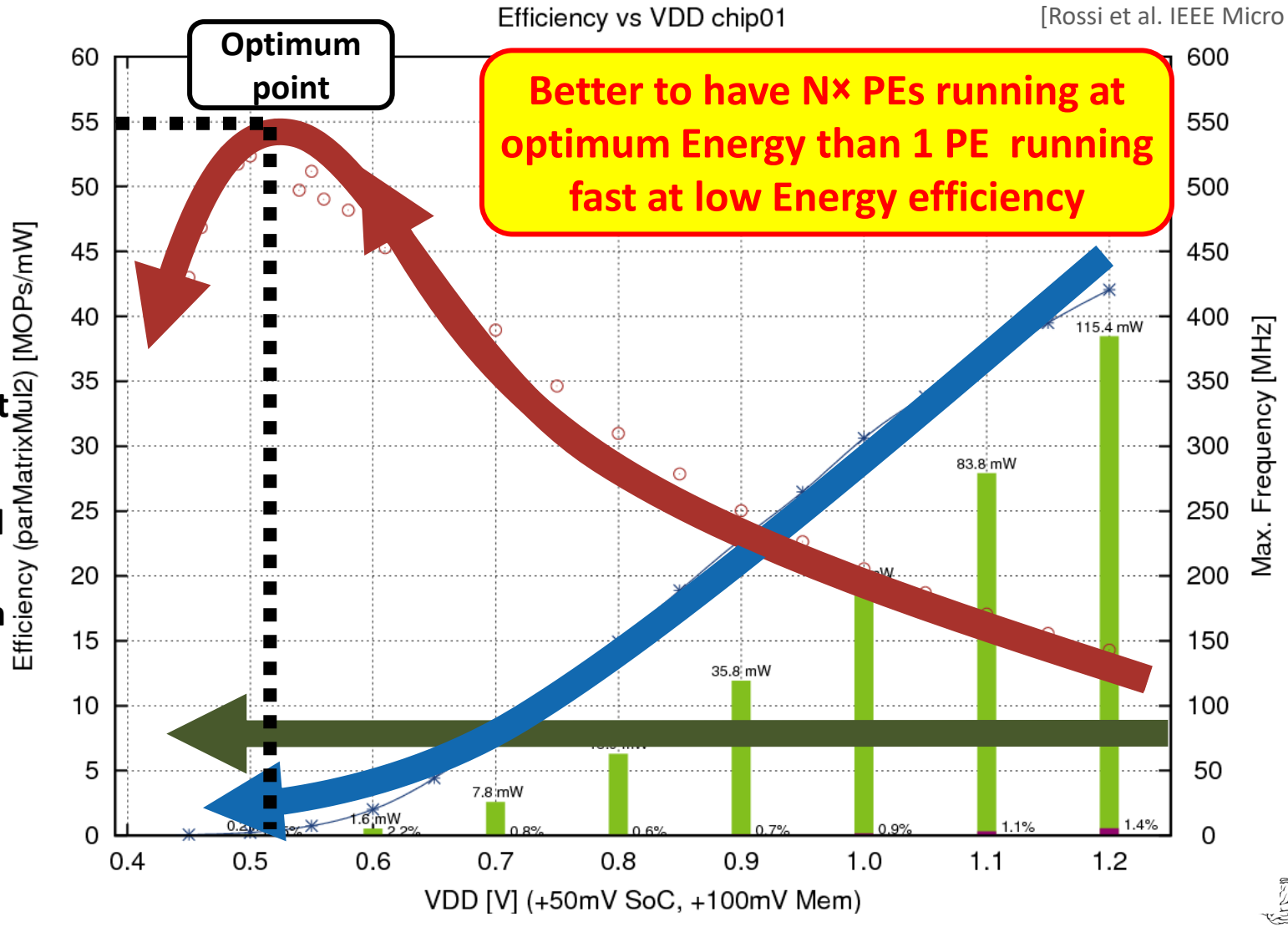


*data from ARM's web

Scaling performance: Parallel, Ultra-Low Power (PULP)

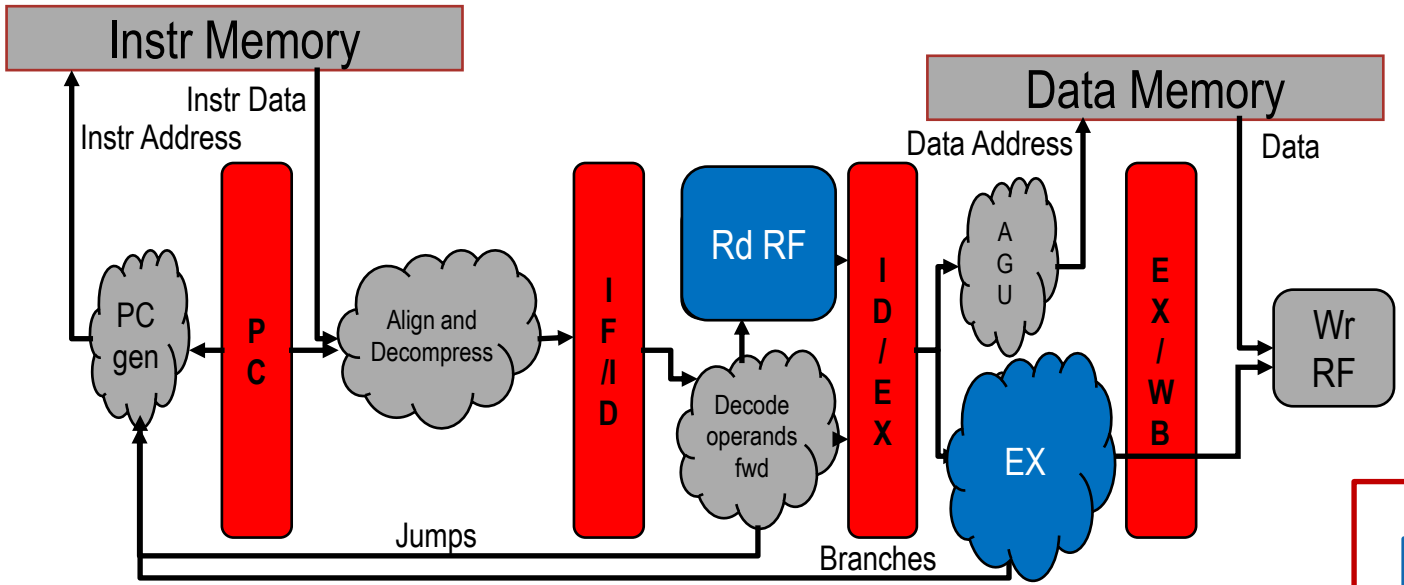
- As VDD decreases, operating speed decreases
- However efficiency increases→ more work done per Joule
- Until leakage effects start to dominate
- Put more units in parallel to get performance up and keep them busy with a parallel workload

If workload is parallel: ML and DSP are parallelizable (embarrassing so)

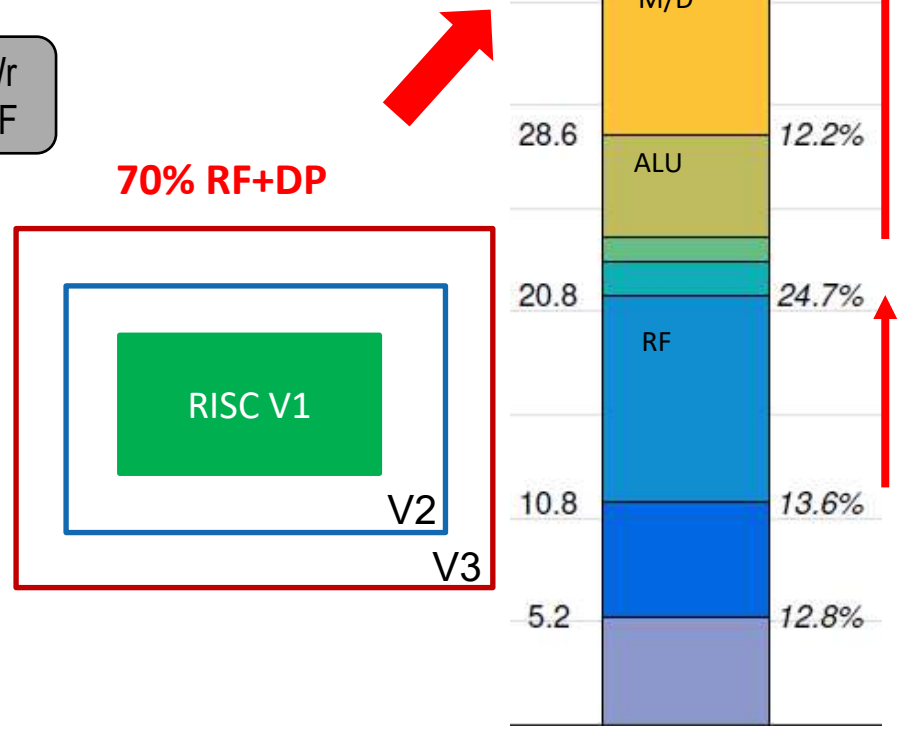


Processor Specialization

3-cycle ALU-OP, 4-cyle MEM-OP→ only IPC loss: LD-use, Branch



[Gautschi et al. TVLSI 2017]



RISC-V® **ISA is extensible *by construction* (great!)**

V1 Baseline RV **(not good for ML)**

V2 Extensions for Data Processing
Data motion (e.g. auto-increment)
Data processing (e.g. MAC)

V3 Domain specific data processing
Narrow bitwidth
HW support for special arithmetic

ISA extension cost 25 kGE → 40 kGE (1.6x), energy efficient if 0.6Texec

Achieving 100% dotp Unit Utilization

8-bit Convolution

HW Loop

LD/ST with post increment

8-bit SIMD sdotp

8-bit sdotp + LD

RV32IMC

N

```
addi   a0,a0,1
addi   t1,t1,1
addi   t3,t3,1
addi   t4,t4,1
lbu    a7,-1(a0)
lbu    a6,-1(t4)
lbu    a5,-1(t3)
lbu    t5,-1(t1)
mul    s1,a7,a6
mul    a7,a7,a5
add    s0,s0,s1
mul    a6,a6,t5
add    t0,t0,a7
mul    a5,a5,t5
add    t2,t2,a6
add    t6,t6,a5
bne    s5,a0,1c000bc
```

RV32IMCXpulp

N/4

```
lp.setup
p.lw  w1, 4(a0!)
p.lw  w2, 4(a1!)
p.lw  x1, 4(a2!)
p.lw  x2, 4(a3!)
pv.sdotsp.b   s1, w1, x1
pv.sdotsp.b   s2, w1, x2
pv.sdotsp.b   s3, w2, x1
pv.sdotsp.b   s4, w2, x2
end
```

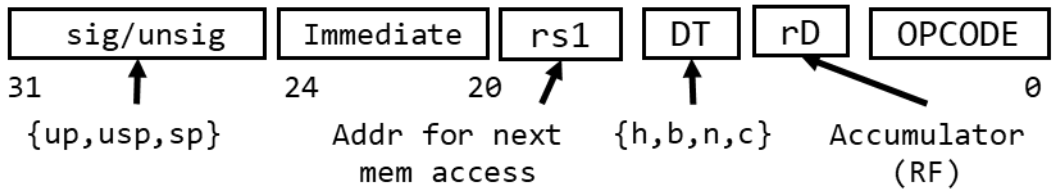
can we remove?

Yes! `dotp+ld`

N/4

```
Init NN-RF (outside of the loop)
lp.setup
pv.nnsdotup.h  s0,ax1,9
pv.nnsdotsp.b  s1, aw2, 0
pv.nnsdotsp.b  s2, aw4, 2
pv.nnsdotsp.b  s3, aw3, 4
pv.nnsdotsp.b  s4, ax1, 14
end
```

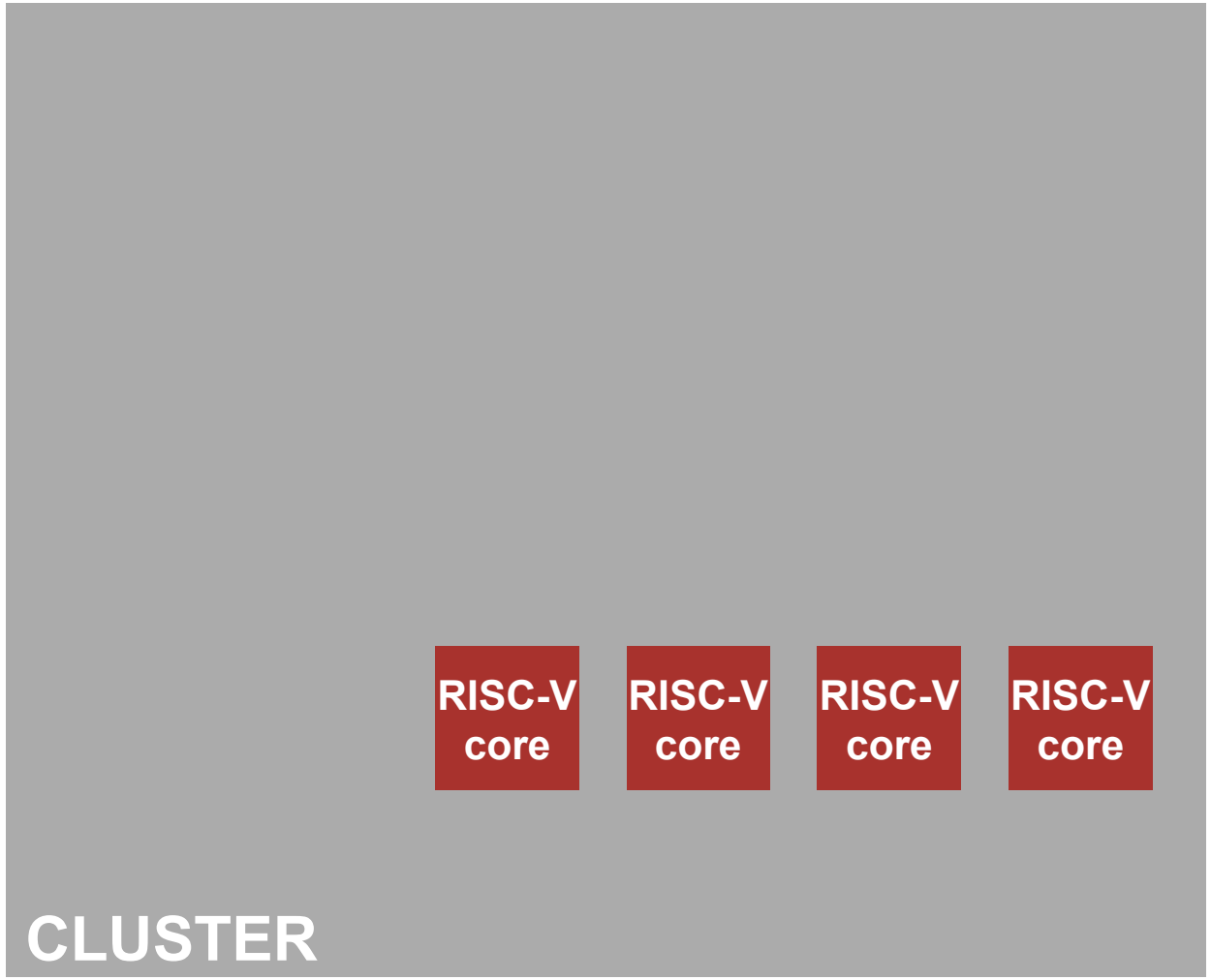
`pv.nnsdot{up,usp,sp}.{h,b,n,c} rD, rs1, Imm`



9x less instructions than RV32IMC

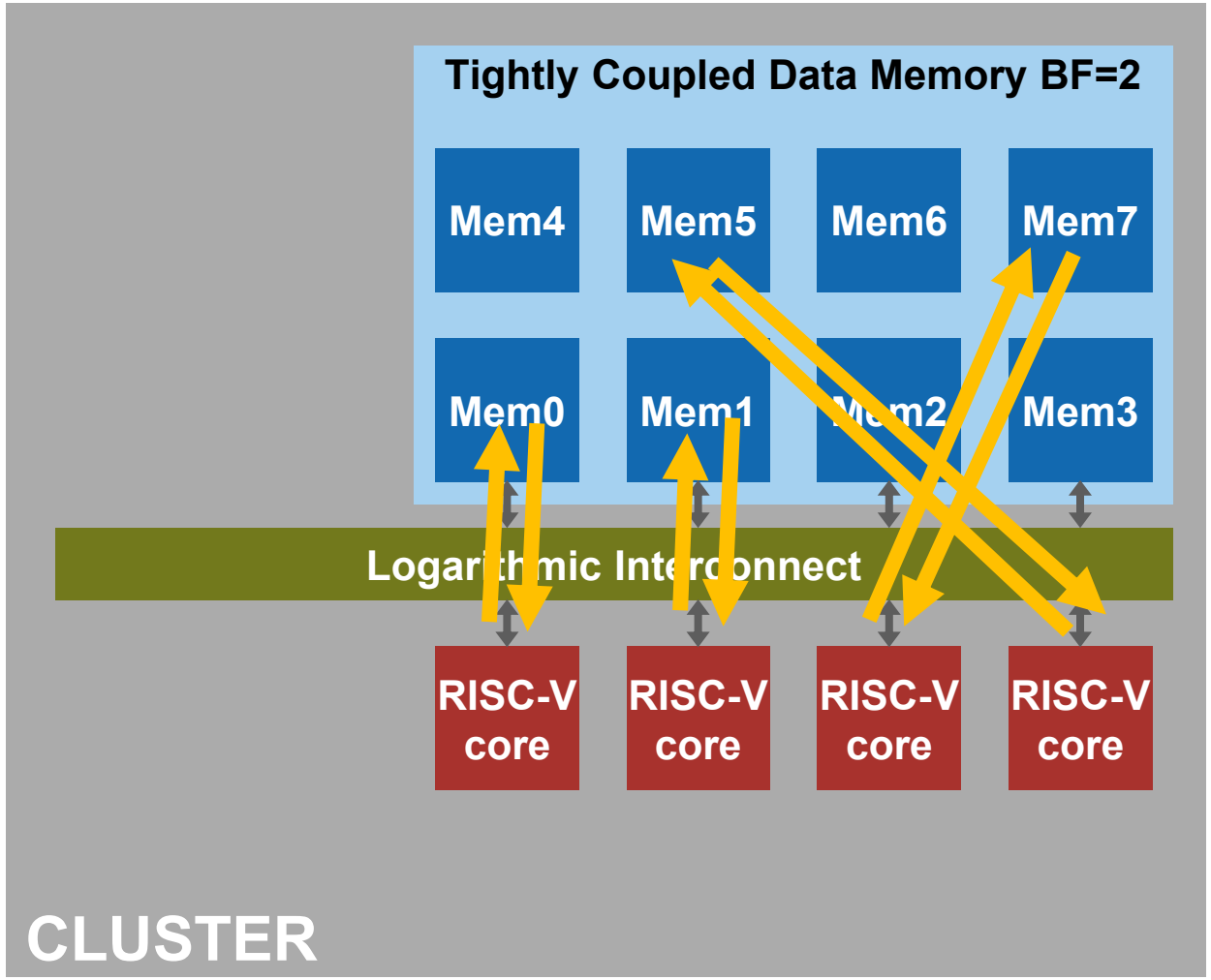
14.5x less instructions at an extra 3% area cost (~600GEs)

Multiple RI5CY Cores (1-16)

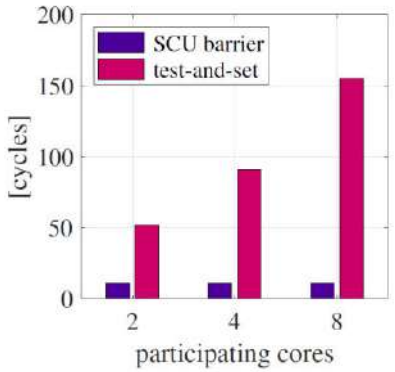


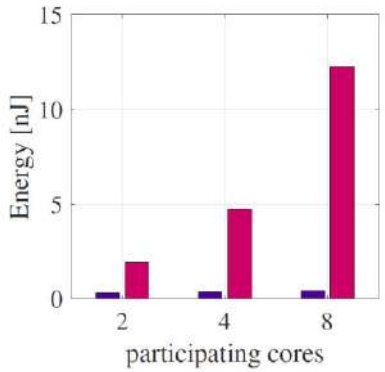
Low-Latency Shared TCDM

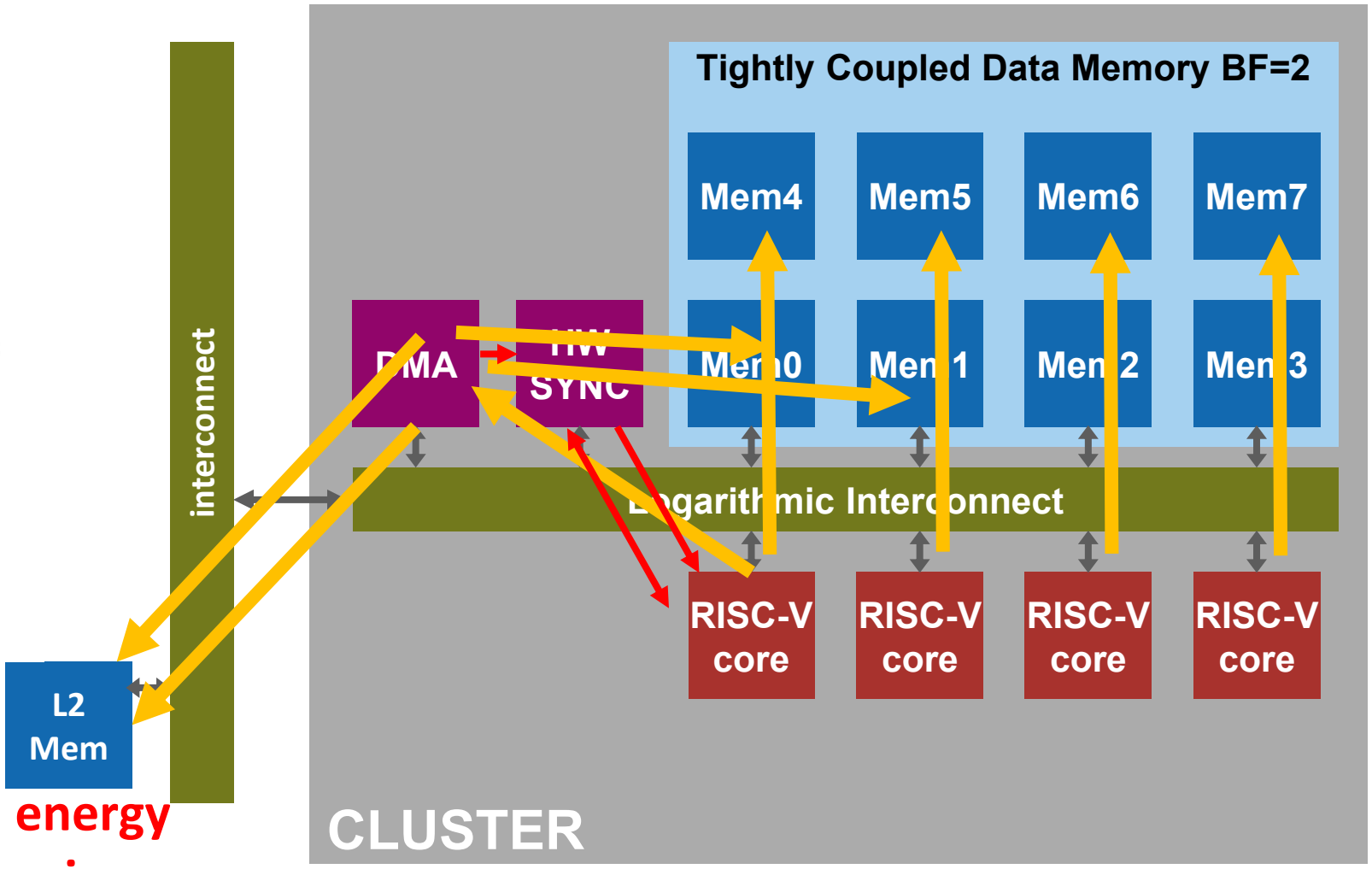
- **Parallel memory access with low contention**
  - Multi-banked, address-interleaved L1
- **Fast interconnect with physical design awareness**
  - Logarithmic depth of combinational switchboxes



Fast synchronization, non-blocking DMA L1-L2 copies





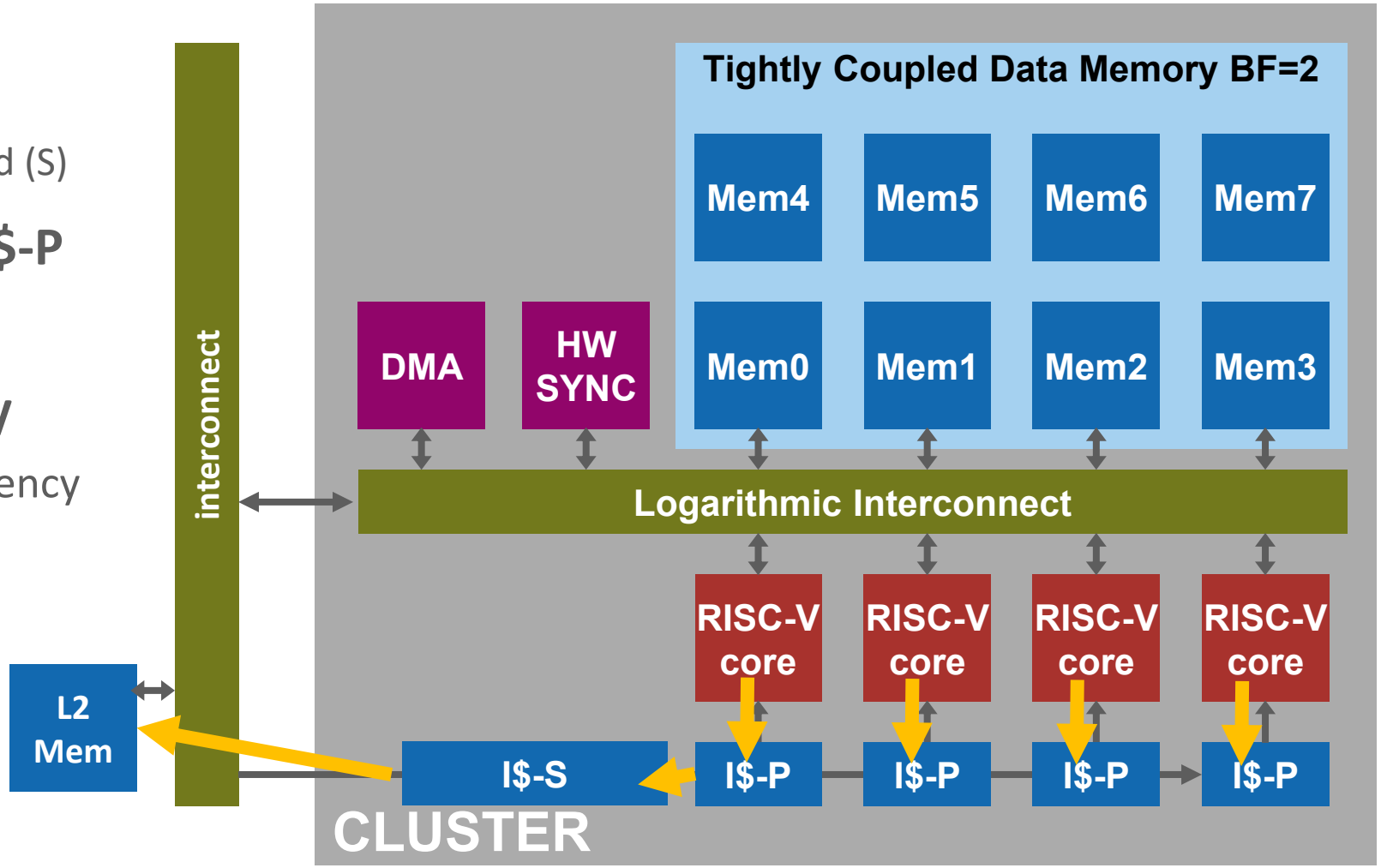


~15x latency and energy reduction for a barrier

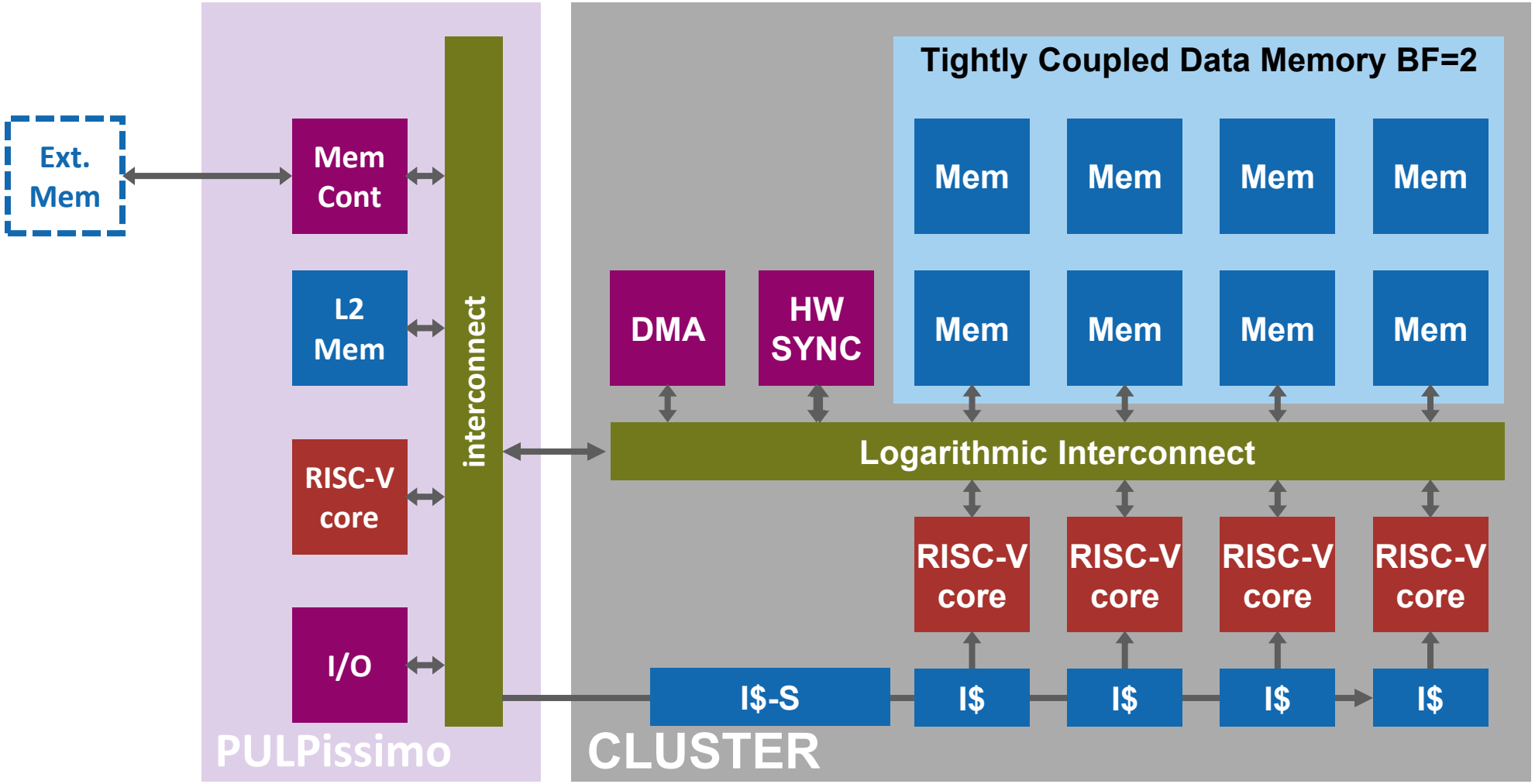
[Glaser TPDS20]

Shared instruction cache with private “loop buffer”

- **Two-level I$**
  - Private (P) + Shared (S)
- **Most IFs from I$-P**
  - Low IF energy
- **I$-S for capacity**
  - Reduces miss latency



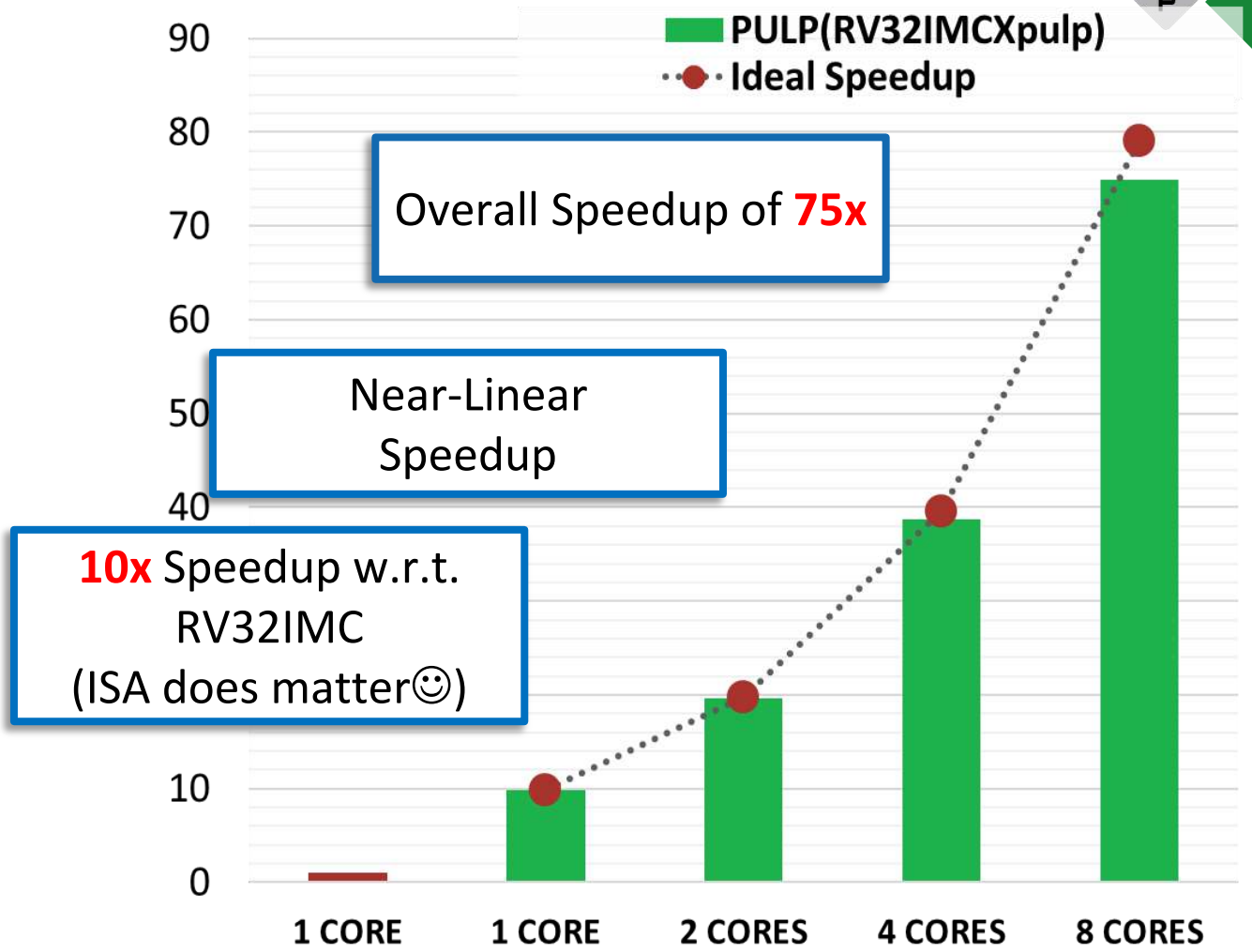
Host for sequential, I/O + Data-Parallel Cluster



Open sourced since 2017: **github.com/pulp-platform/pulp**

Combining ISA extension + Efficient parallel execution

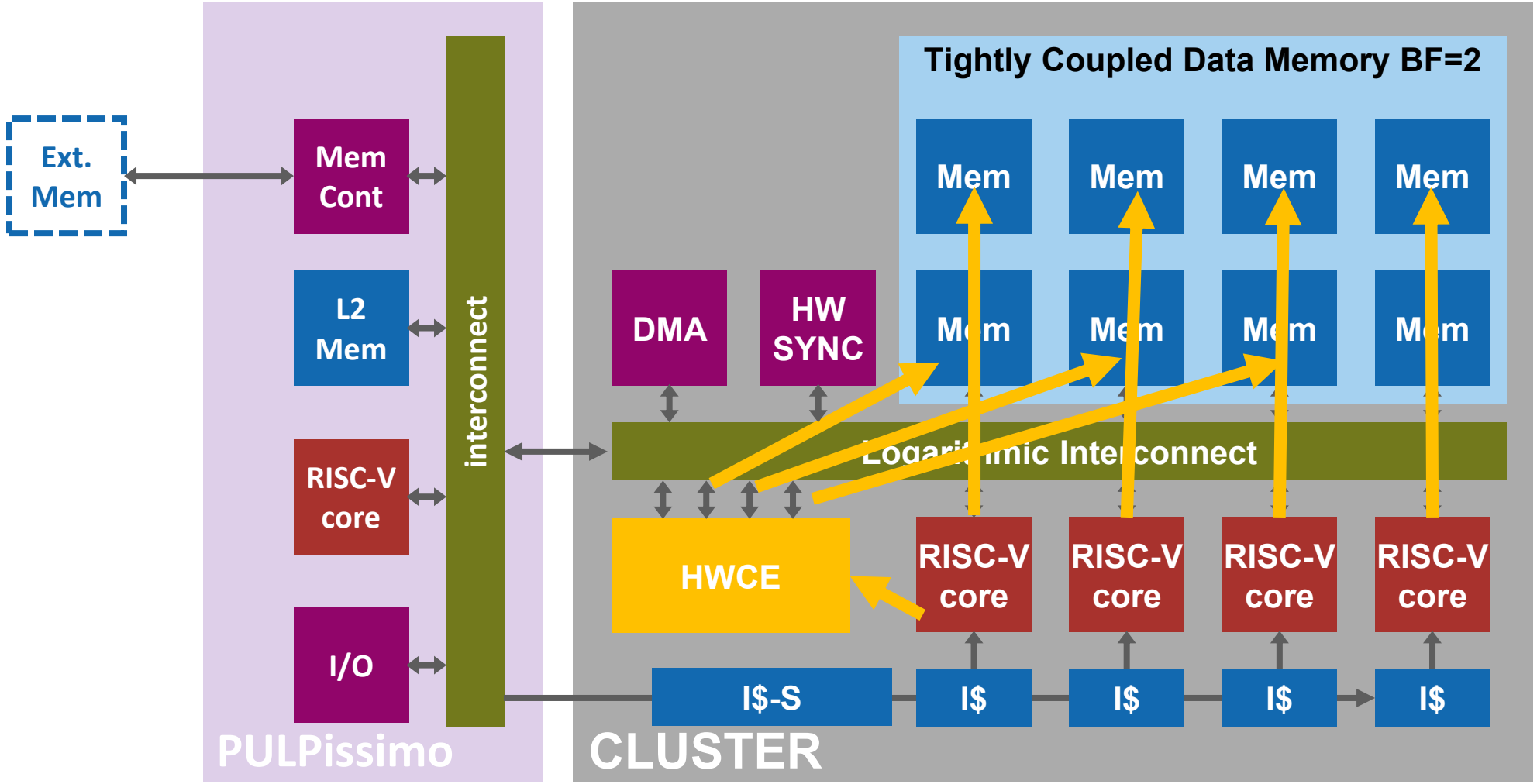
- **8-bit convolution**
  - Open source DNN library
- **10x through xPULP**
  - Extensions bring real speedup
- **Near-linear speedup**
  - Scales well for regular workloads
- **75x overall gain**
- **7-8 GMACs**
  - 250MHz
  - 4 MAC/Cycle (8bit)
  - 8 Cores



[Garofalo et al. Philos. Trans. R. Soc 20]

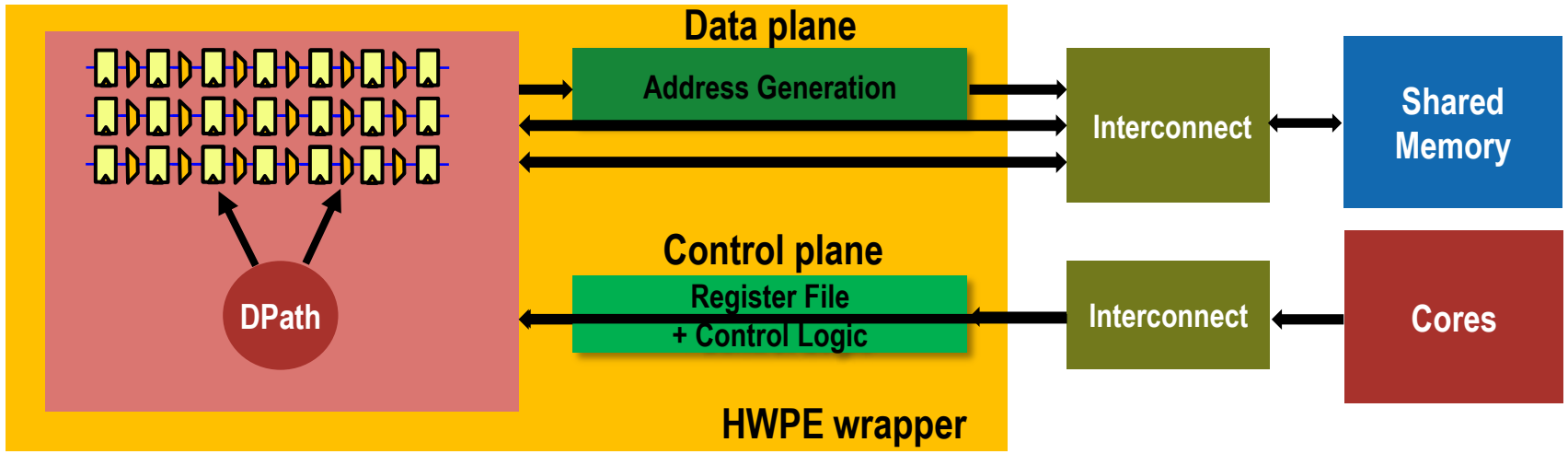
More GOPS, Less Power?

What's next? Tightly-coupled HW Compute Engine



Acceleration with flexibility: zero-copy HW-SW cooperation

Hardware Processing Engines (HWPEs)

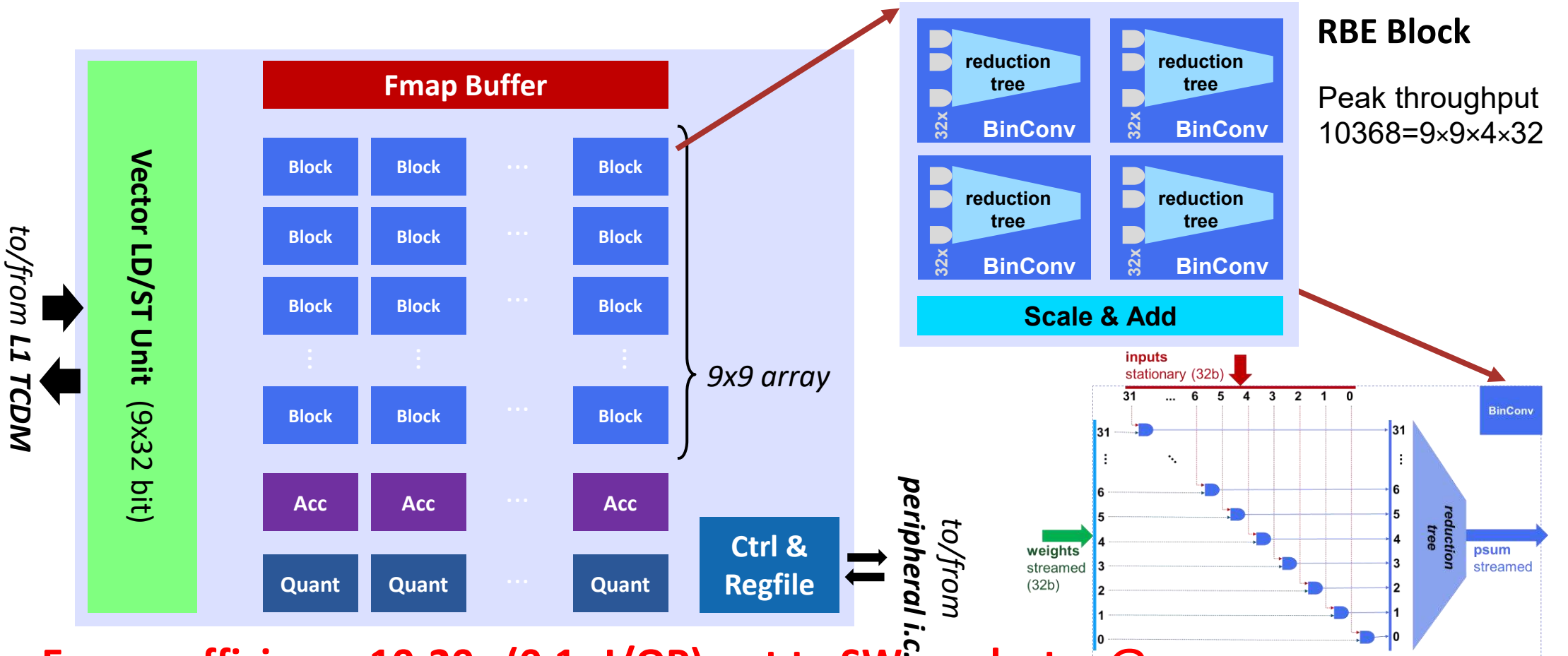


HWPE efficiency ($\frac{MAC}{A(mm^2),E(J),W(bps)}$) vs. optimized RISC-V core

1. Dedicated control (no I-fetch) with shadow registers (overlapped config-exec)
2. Specialized high-BW interco into L1 (on data-plane)
3. Specialized datapath: supporting configurable & aggressive quantization

Reconfigurable Binary Engine

$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j (\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in}))\right)$$

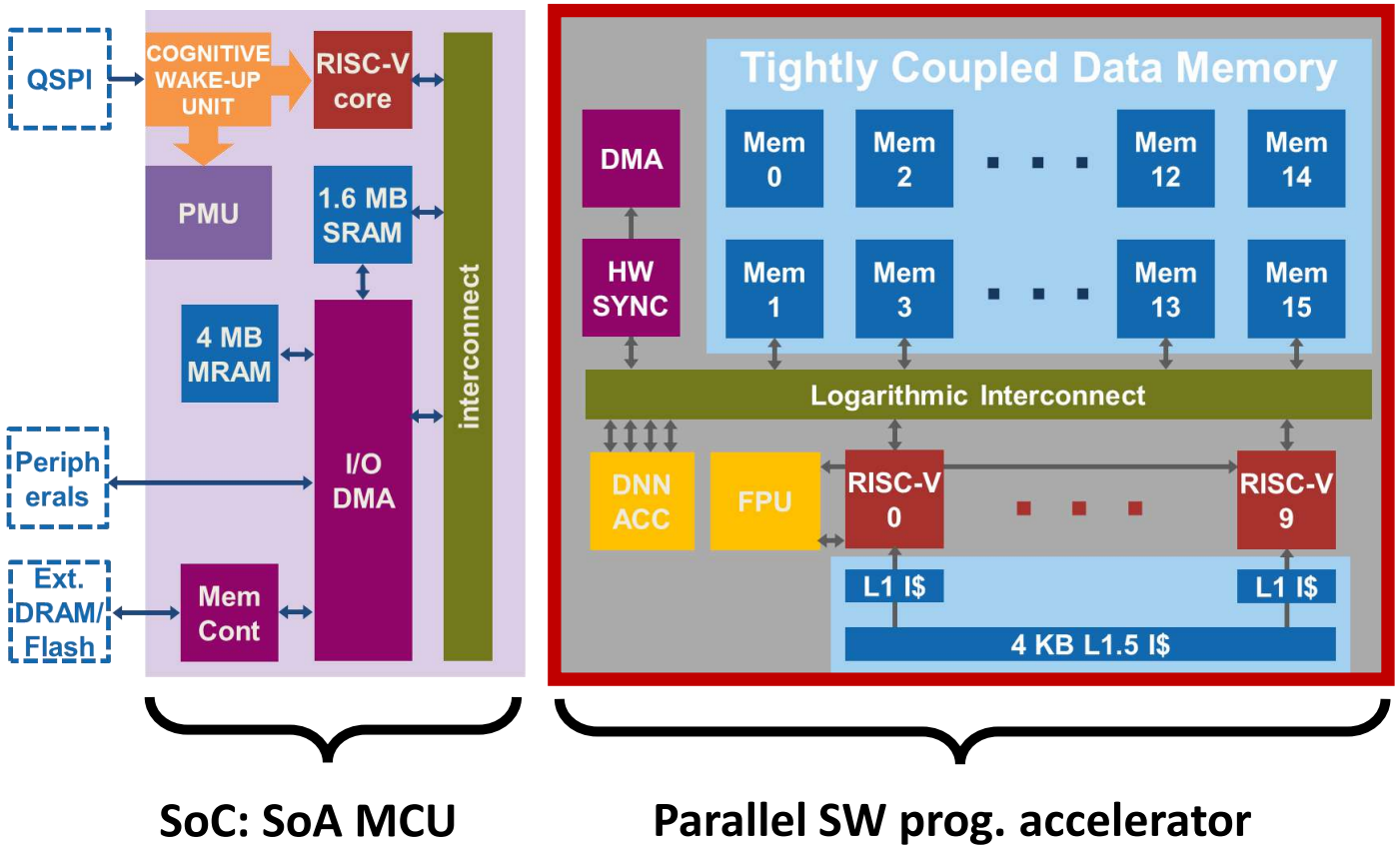


Energy efficiency 10-20x (0.1pJ/OP) wrt to SW on cluster @same accuracy

All together in VEGA: Extreme Edge IoT Processor

[Rossi et al. ISSCC21]

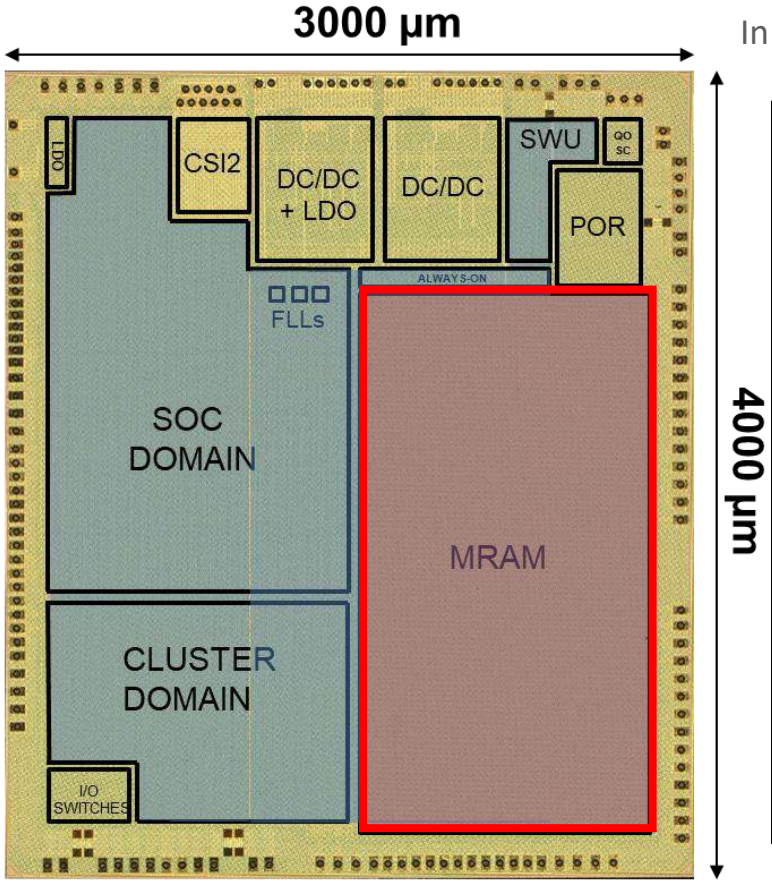
- RISC-V cluster (8cores +1) 614GOPS/W @ 7.6GOPS (8bit DNNs), 79GFLOPS/W @ 1GFLOP (32bit FP appl)
- Multi-precision HWCE(4b/8b/16b) 3×3×3 MACs with normalization / activation: 32.2GOPS and 1.3TOPS/W (8bit)
- 1.7 µW cognitive unit for autonomous wake-up from retentive sleep mode



In cooperation with GREENWAVES TECHNOLOGIES

All together in VEGA: Extreme Edge IoT Processor

- RISC-V cluster (8cores +1) 614GOPS/W @ 7.6GOPS (8bit DNNs), 79GFLOPS/W @ 1GFLOP (32bit FP appl)
- Multi-precision HWCE(4b/8b/16b) 3×3×3 MACs with normalization / activation: 32.2GOPS and 1.3TOPS/W (8bit)
- 1.7 µW cognitive unit for autonomous wake-up from retentive sleep mode
- **Fully-on chip DNN inference with 4MB MRAM (high-density NVM with good scaling)**

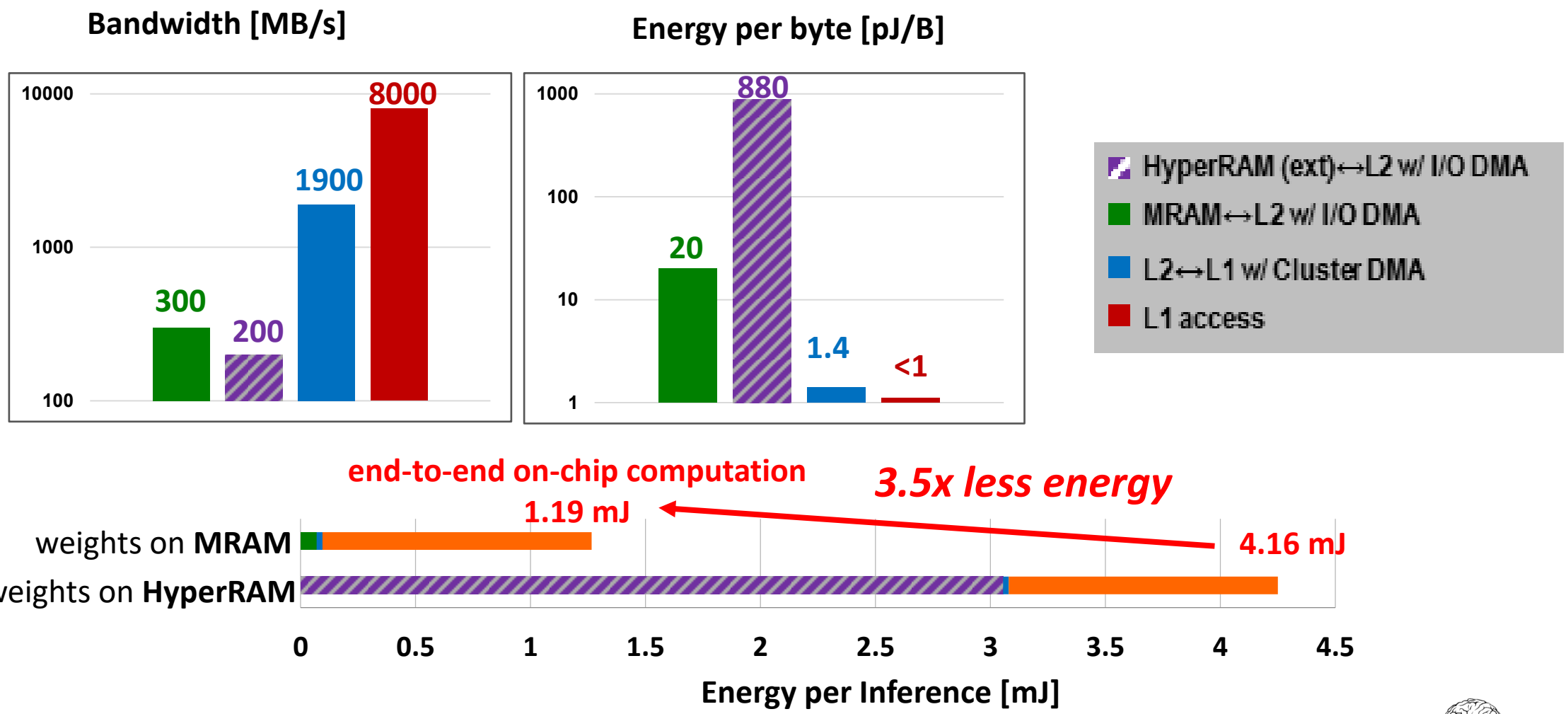


[D. Rossi, ISSCC21]

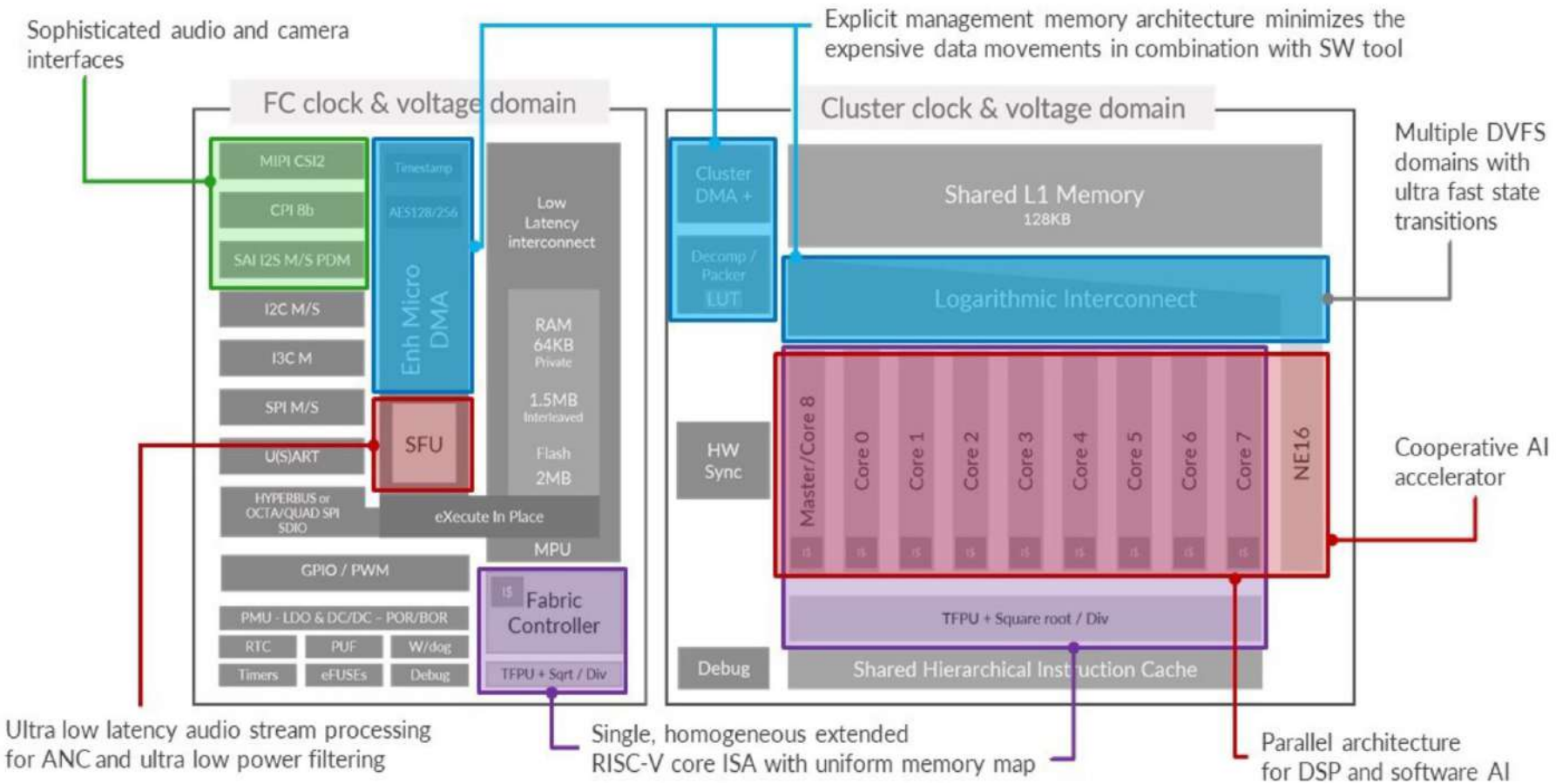
In cooperation with GREENWAVES TECHNOLOGIES

Technology	22nm FDSOI
Chip Area	12mm^2
SRAM	1.7 MB
MRAM	4 MB
VDD range	0.5V - 0.8V
VBB range	0V - 1.1V
Fr. Range	32 kHz - 450 MHz
Pow. Range	1.7 µW - 49.4 mW

Full DNN Energy (MobileNetV2) on Vega



PULP→GAP8, VEGA→GAP9



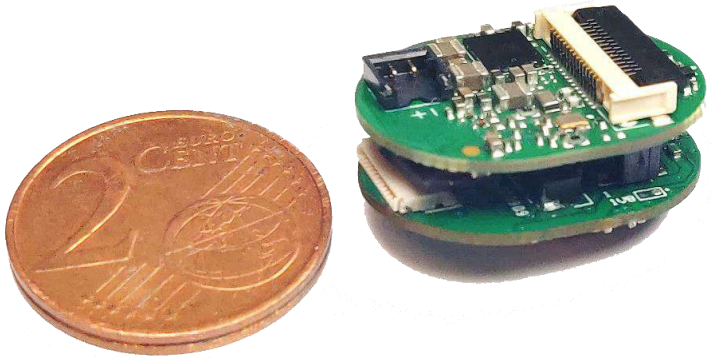
Respectively 85% and 65% of GAP8 and GAP9 are based on open-source IPs

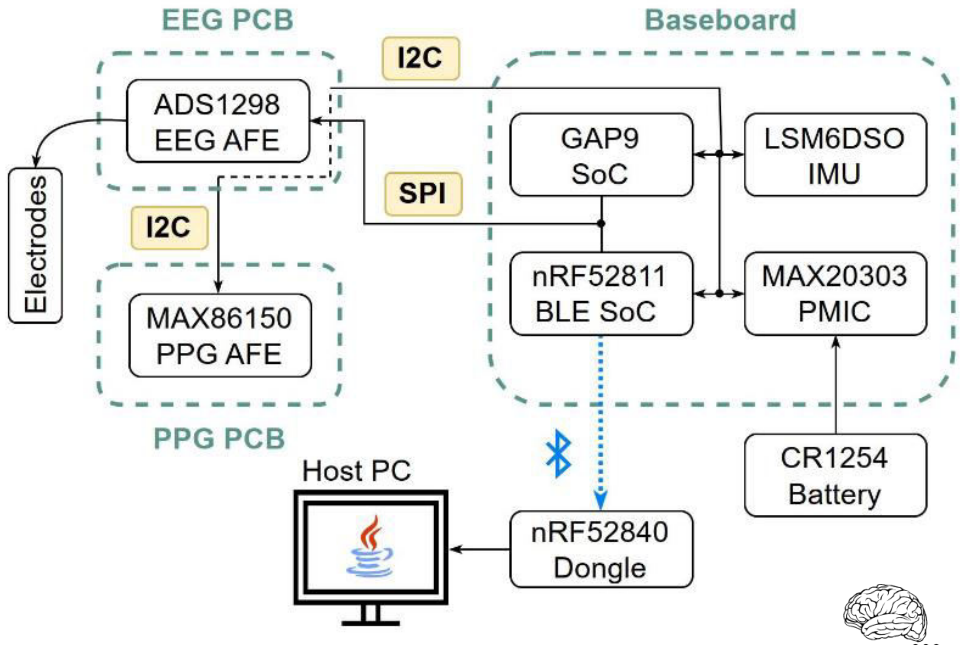
Wearable EEG needs compute…

plus body interface & wireless communication

BioGAP: the ultimate wearable computing platform

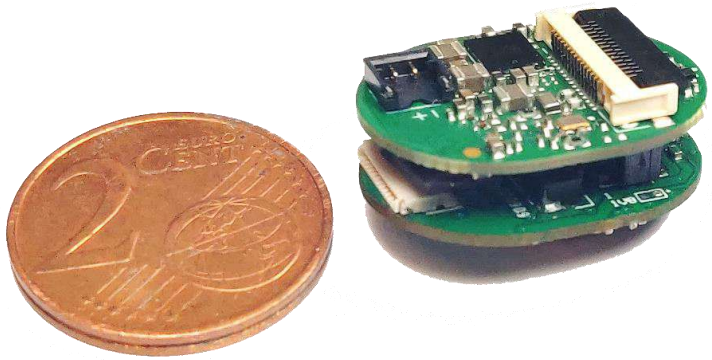
- PULP-based computing platform (GAP9)
- Nordic nRF52 for BLE connectivity
- Can be flexibly connected to a large variety of sensor interfaces
  - EEG shield for measurement of biopotentials
  - PPG board
- Sub-20mW power consumption

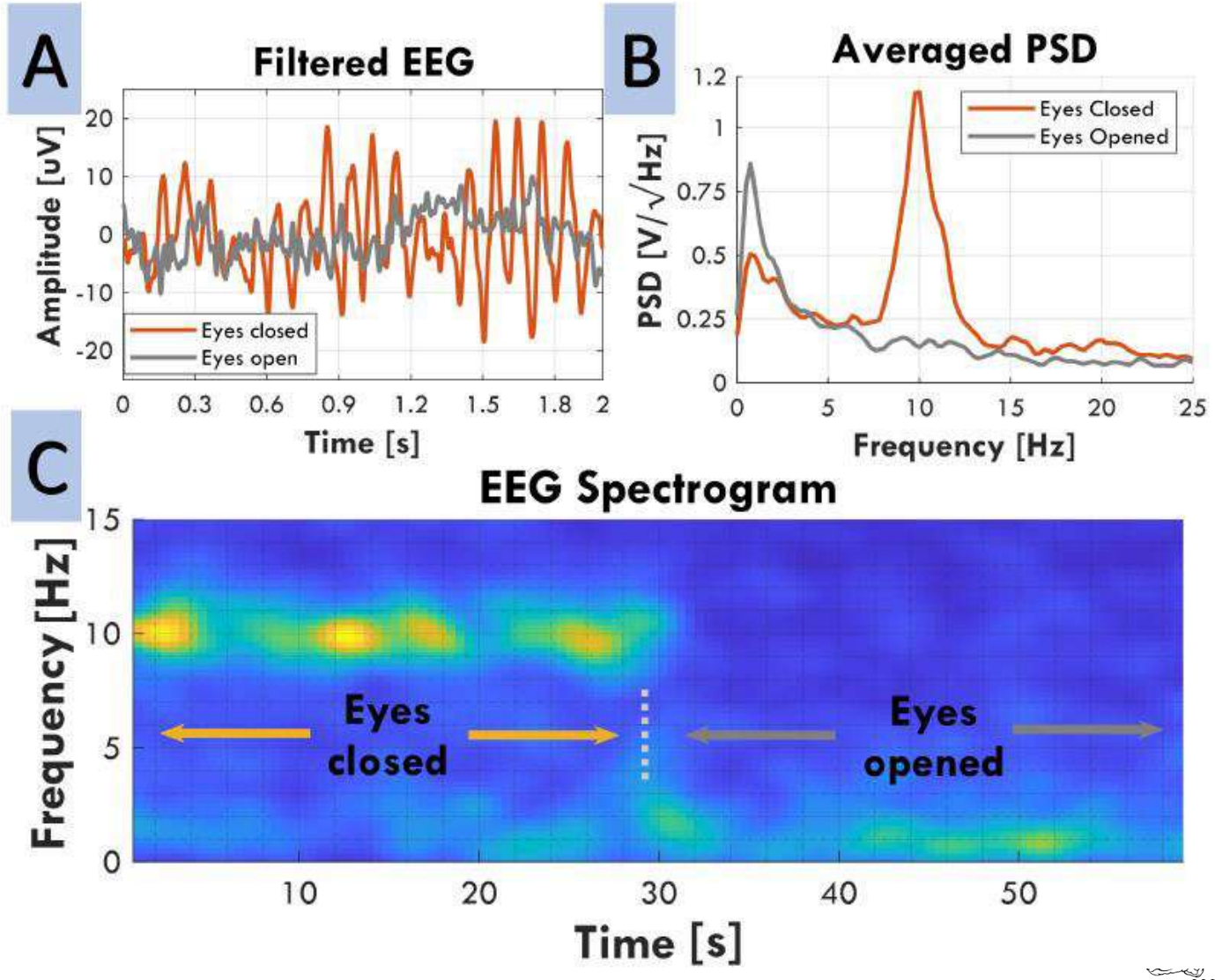




BioGAP: the ultimate wearable computing platform

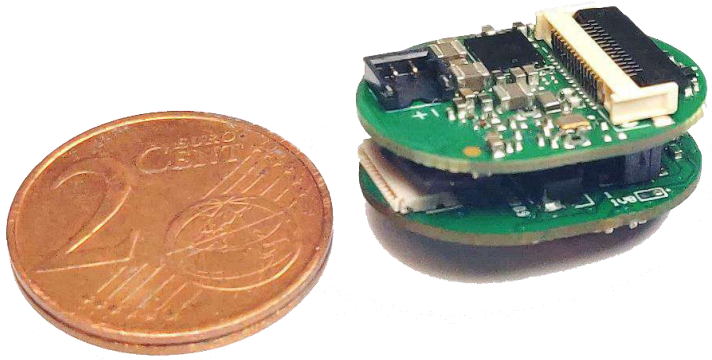
- Validation on alpha waves

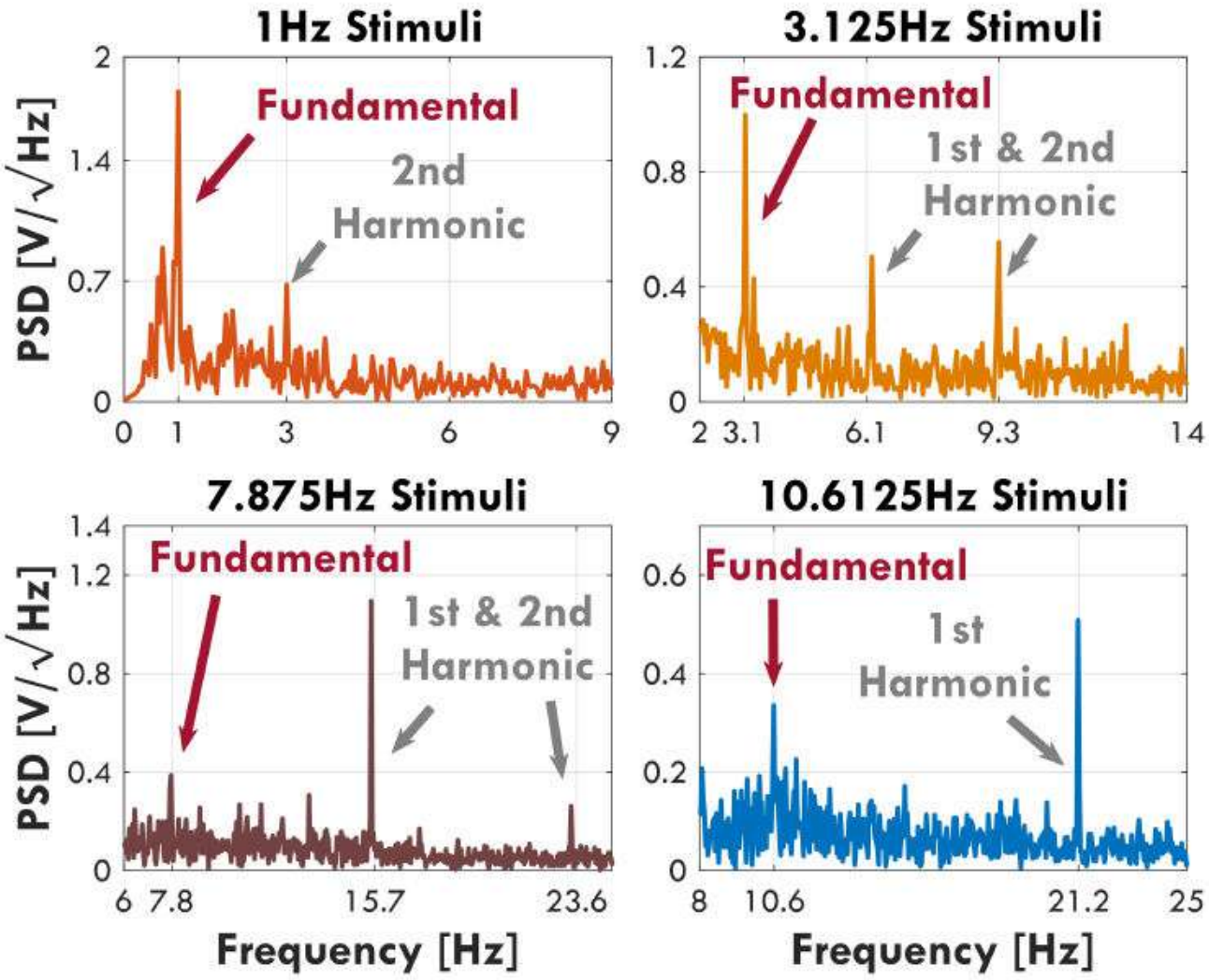




BioGAP: the ultimate wearable computing platform

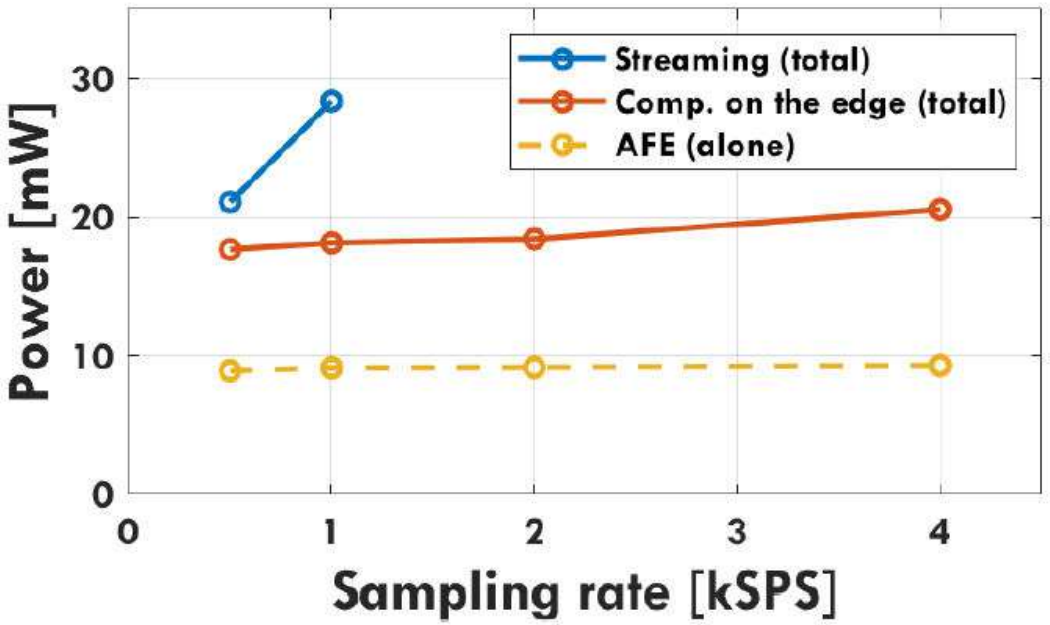
- Validation on SSVEP

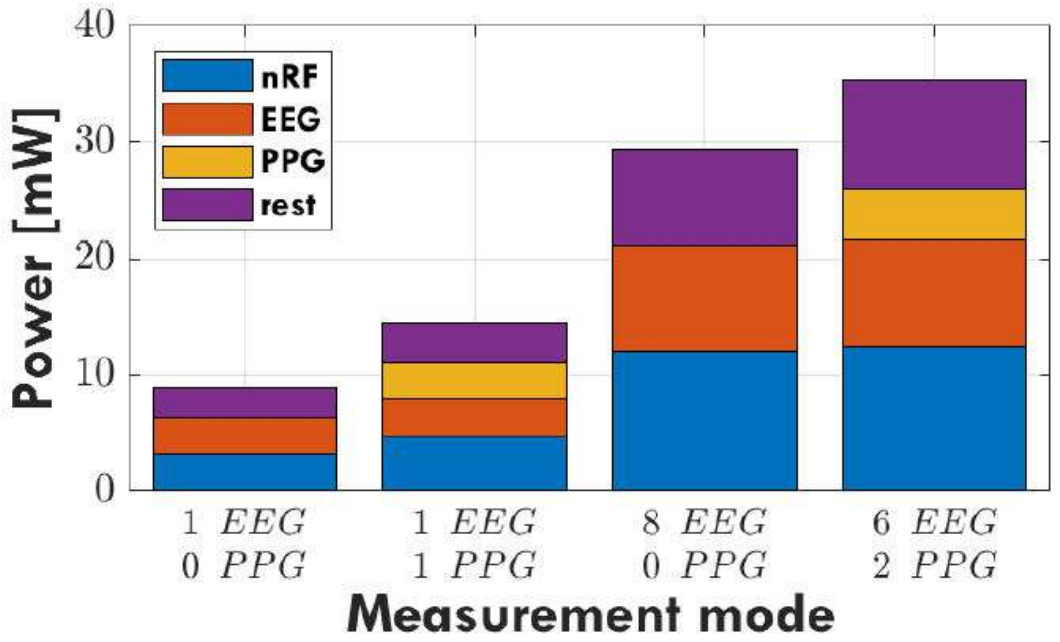




BioGAP: the ultimate wearable computing platform

- Power performance: Computing onboard enables higher sampling rates





Headbands for Epilepsy Monitoring Units and BCIs

- One version (white) for epilepsy monitoring units
  - Brush electrodes to enable measurements in the presence of bandages

DATWYLER

- One version (black) for ambulatory and consumer BCI applications
  - Spider electrodes, more comfortable

DATWYLER

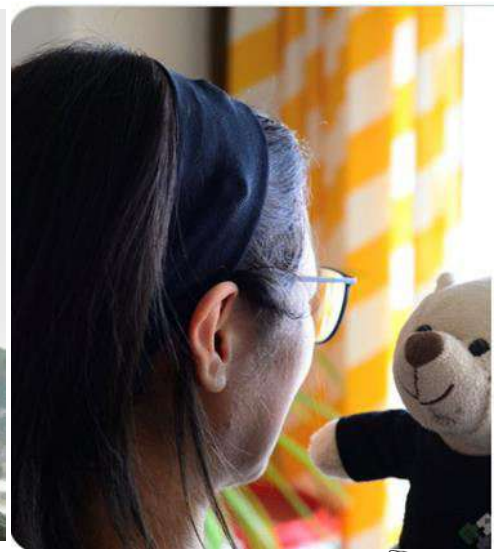
Headbands for Epilepsy Monitoring Units and BCIs

- One version (white) for epilepsy monitoring units
  - Brush electrodes to enable measurements in the presence of bandages

DATWYLER

- One version (black) for ambulatory and consumer BCI applications
  - Spider electrodes, more comfortable

DATWYLER

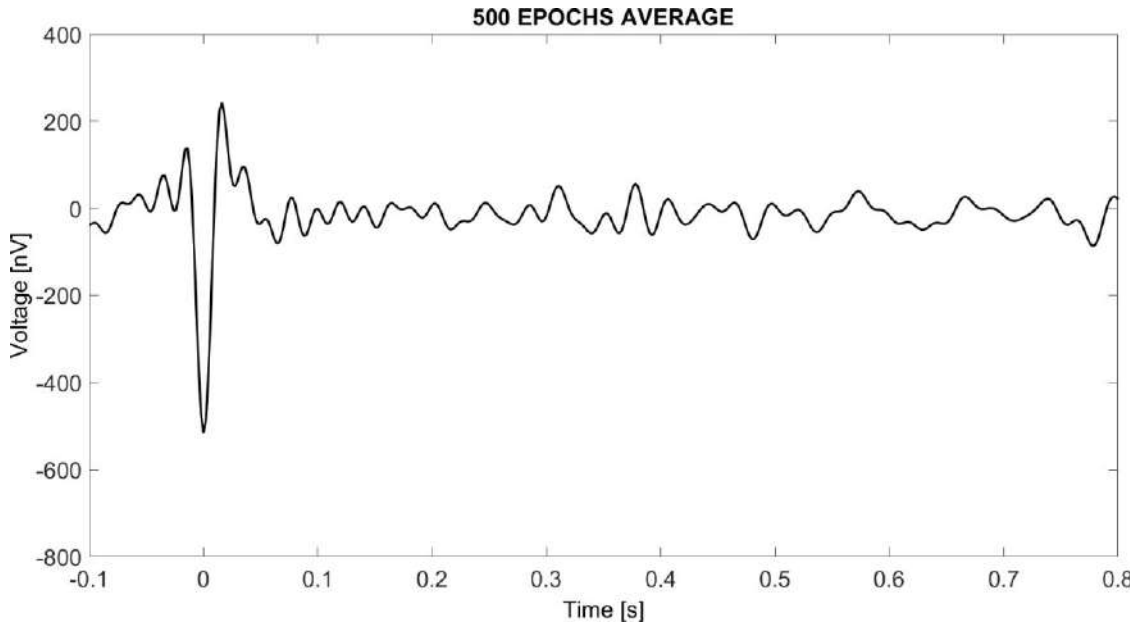


Toward ears EEG

- In-ear electrode:
  - Dätwyler Holding Inc. (derived from SoftPulseTM family)
  - Conductive elastomer
  - Silver/silver-chloride coating to optimize contact with the skin
  - Snap connector interface w/ acquisition PCB
- Reference and bias electrodes:
  - 4mm neodymium magnets
  - Gold coated
  - Allowing for flexible positioning on the ear (inner/outer side of ear scapha)

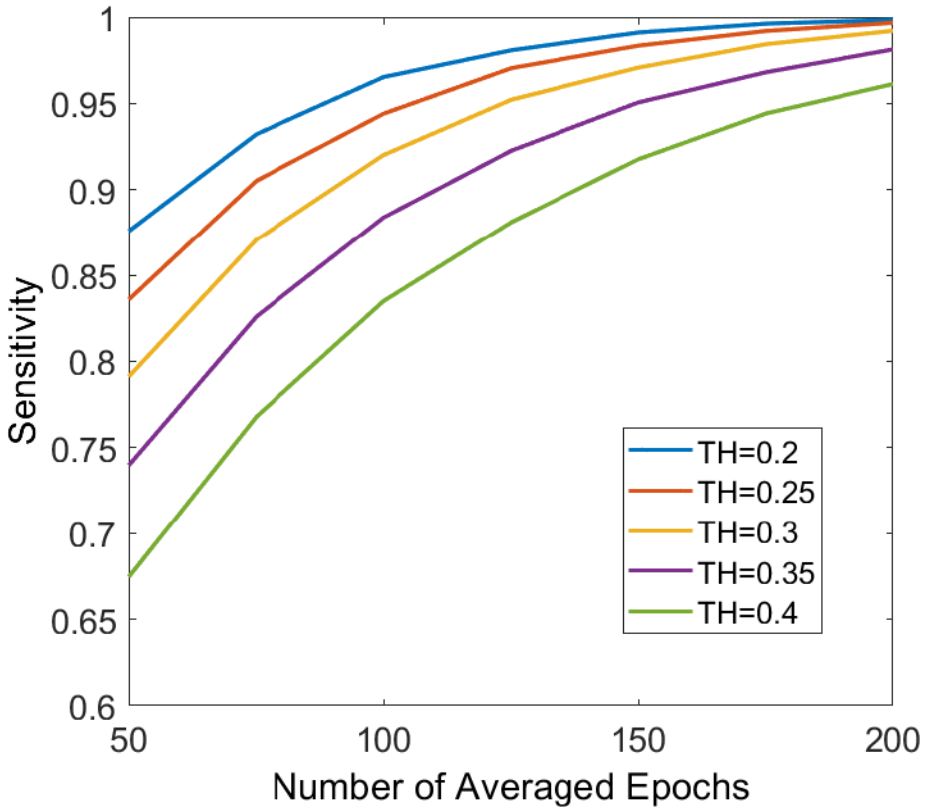
Application: auditory stimuli

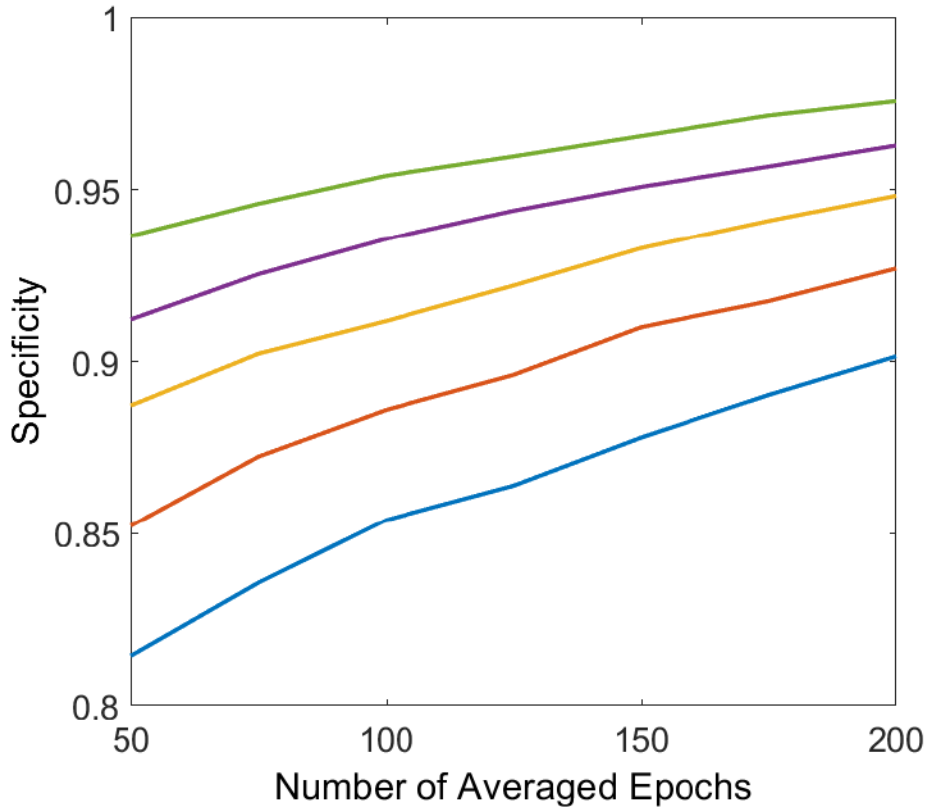
- Time domain (for low computational complexity)
- Stimulus
  - Single 50ms Gaussian noise pulse
  - Random inter stimulus interval (ISI)
    (500 ms + random jitter [0, 200]ms)
- Filtering
  - Band-pass 8-48 Hz + Notch 50 Hz
  - IIR (for reduced complexity)
- Windowing centered on the peak + averaging
- Time-domain correlation to template + thresholding
- Template response: averaging 500 epochs with average ISI = 600 ms
- Control: 5 minutes resting state (speakers turned off)



Few epochs are enough for >80% sensitivity/specificity

- Small number of epochs (50) is enough to obtain both sensitivity and specificity > 80%
- Tradeoff between #epochs, threshold, correlation window to penalize sensitivity or specificity





TH = correlation threshold

Results for windows=60 samples

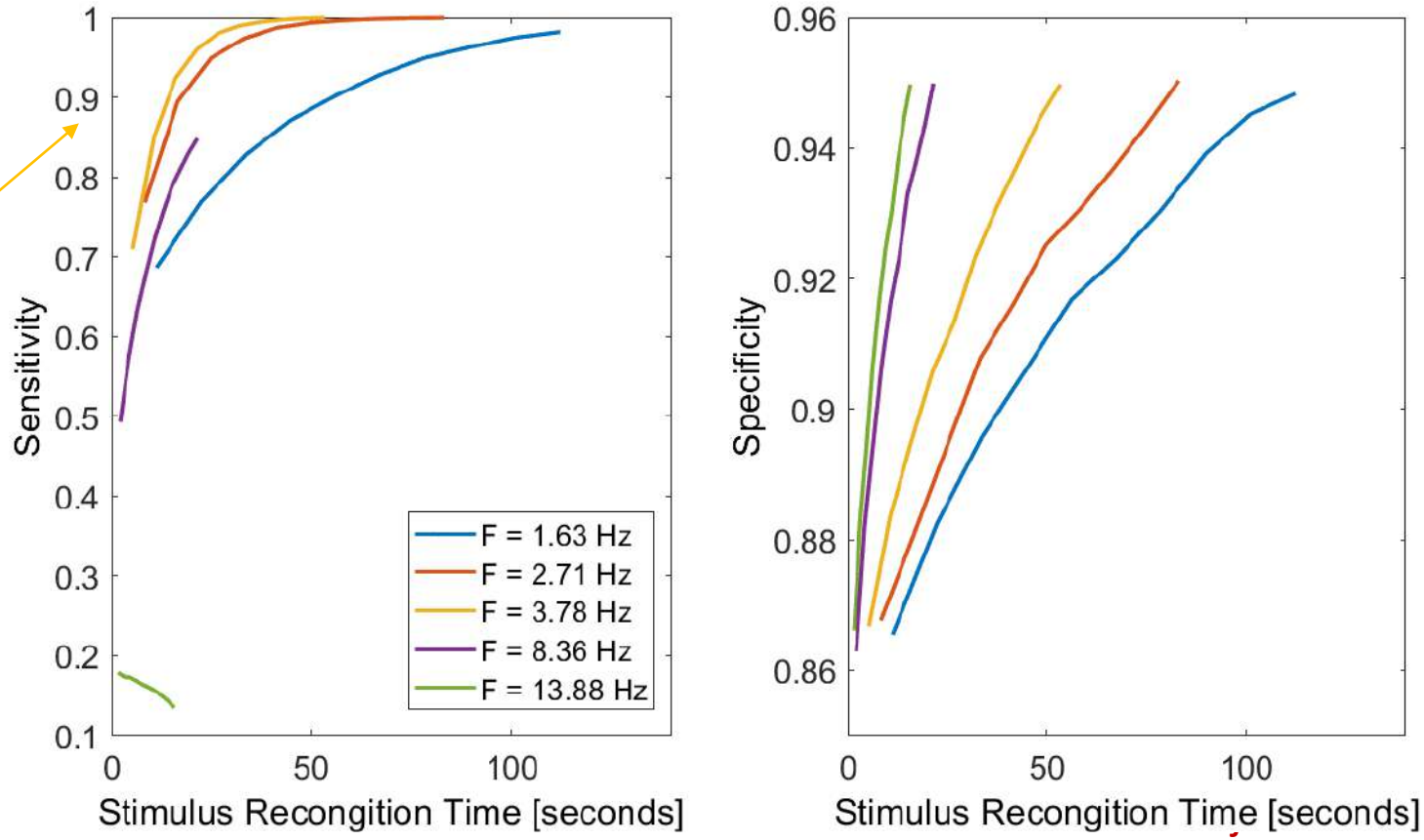
8 subjects

[Guermandi et al., Proc. EMBC, 2022]

Very fast detection time

8 subjects

- Results for different stimulus repetition rates
- Threshold = 0.275
- Correlation window = 60 samples
- Max speed for sensitivity at F≈4 Hz
- Very fast detection time (SoA systems use 100s trials and 500s stimulation times)



[Guermandi et al., Proc. EMBC, 2022]

earEEG present achievements and future

- ear-EEG system with on-board processing completely embedded in an earbud-like form factor
- Sensitivity and specificity >80% obtained with a small number of epochs (50)
- Low-power (1.3 mW) and almost one-month of battery lifetime
- Demonstrated the potential of the proposed system for objective hearing threshold estimation
- The device could be integrated into standard earbuds or hearing aid devices

- WIP:
  - Integration of BioGAP in an earbud form factor for enhanced computing capabilities

Thank You!